Linguistic determinants of English personal name choice

Stephanie S Shih
Departments of Linguistics
Stanford University
University of California, Berkeley

LSA Annual Meeting
Portland, Oregon
7 January 2012

slides: http://www.stanford.edu/~stephsus/shih-LSA2012.pdf

## Personal name choice

- Personal name = Forename + Surname
  *choice* *fixed*

- Previously identified factors in name choice
  - ethnic, cultural, religious, socioeconomic, educational background and communities (Bloothooft and Groot 2008; Mateos and Tucker 2008; Barry 2010; Bloothooft and Onland 2011; a.o.)
  - naming trends, popularity, frequency (Tucker 2001; a.o.)
  - sound symbolism (Whissell 2001)

  → external linguistic factors

- Forenames and surnames are usually studied independently.

## Personal name choice

- forename + surname = integrated unit
- phrasal stress: JOHN + SMITH → john SMITH
- popular baby-naming advice:
  "A full name is like a little line of poetry…. You may choose a name you love, only to test it out with your surname and find it falls flat." (Wattenberg 2005)

## Personal name choice

Phonological considerations from baby-naming advice:

- "The baby first name's rhythm should match the last name…. Say the first, middle, and last name several times to test the rhythm. Say the first and last name together, too." (www.circleofmoms.com)
- "Look carefully where the end of one name meets the beginning of another. Jonas Sanders will be heard as Jonah Sanders or Jonas Anders." (Wattenberg 2005: 4)

## Personal name choice

- Do internal linguistic, phonological factors determine personal name choice across first and last name pairs?

- Phonological factors affect other linguistic choices (e.g., word order, construction choices).

## Phonology in linguistic choice

- Avoidance of adjacent sibilant segments affects English genitive construction choice. (Menn and MacWhinney 1984; Zwicky 1987; Hinrichs and Szmrecsányi 2007; et seq.)

  *the wheel of the bus > the bus's wheel*
  *the bell of the church > the church's bell*

## Phonology in linguistic choice

- Rhythmic well-formedness preferences affect word order and construction choices. (McDonald et al. 1993; Benor and Levy 2006; Shih et al., to appear; a.o.)
  - lapse avoidance:
    *surPRISE and SIN > SIN and surPRISE*
    *the CHILdren's VOIces > the VOIces of the CHILdren*
  - clash avoidance:
    *the SMELL of WHEAT > WHEAT'S SMELL*

## Other potential phonological factors

- Alliteration
  - processing, production, and perceptual benefits (Boers and Lindstromberg 2005; Lindstromberg and Boers 2008; a.o.)
  - Numerous phonological processes cross-linguistically promote segmental agreement
    - alliteration and rhyme in linguistic art forms
    - long distance consonant agreement and other harmony patterns (Zuraw 2002; Rose and Walker 2004; Adams 2010; a.o.)

## Personal name choice

- Do the same phonological factors that affect other linguistic choices also determine personal name choice across first and last name pairs?

Phonological determinants investigated:

- Alliteration
- Avoidance of adjacent identical segments (OCP)
- Rhythmic well-formedness preferences

➢ Phonological factors active in other linguistic choices (e.g., word order) are also active in personal name choice.
→ Speakers utilize the same preferences in choosing names as they do in other linguistic processes.

## Data

- Obstacles to large-scale personal name studies
(cf. Tucker 2001)
  - digitization limitations
  - proprietary information
  - privacy concerns (SSA waits 100 years before releasing full name pairs)

## **Data:** the facebook names corpus

- All publicly available and searchable profiles from www.facebook.com (Bowes 2010) = 171 million personal names (100 million unique)
- The facebook names corpus = 41 million personal names (3.3 million unique)

  Excludes:
  - personal names with only one instance
  - names with more than two orthographic words
  - names in which one name contained only one orthographic letter
  - business names (e.g., *Rainforest Café*)
  - obvious nicknames, aliases, and fictional characters (e.g., *Lord Voldemort*)
  - names not present in Unisyn lexicon (Fitt 2001) ~ non-English names (e.g., *Rajesh*)

## Corpus and Methodology

- Most popular names

| | |
|---|---|
| *John Smith* | 17204 |
| *David Smith* | 7440 |
| *Michael Smith* | 7200 |

- Automatic stress and segmental annotations from the American English Unisyn lexicon (Fitt 2001; Shih 2011 supplement)
- Methodology
  - 3 phonological factors investigated
  - 2 control factors
  - Linear and logistic regression

## **Factor:** Alliteration

- Prediction: All else being equal, speakers will choose alliterative name pairs.

| | | |
|---|---|---|
| e.g., | *Sarah Smith* | 5039 instances |
| | *Steve Smith* | 4316 |
| | *James Johnson* | 3392 |

- Operationalizing alliteration
  - identical word-initial consonants
  - all vowel-initial words alliterate

## **Factor:** Adjacent identity avoidance

Adjacent sibilants [s, z, ʃ, tʃ, ʒ, dʒ]

- Prediction: All else being equal, speakers will avoid name pairs with adjacent sibilant sounds across the first and last name boundary.

| | | |
|---|---|---|
| e.g., | *Charles Smith* | 1587 instances |
| | *Josh Sanders* | 256 |

## **Factor:** Adjacent identity avoidance

Adjacent identical segments (OCP)

- Prediction: All else being equal, speakers will avoid name pairs with adjacent identical segments across the first and last name boundary.

| | | |
|---|---|---|
| e.g., | *Michael Lee* | 2540 instances |
| | *Michelle Lee* | 2003 |
| | *Robert Taylor* | 1889 |

- Operationalizing OCP
  - identical forename-final and surname-initial consonants
  - all vowels considered identical

## **Factor:** Rhythm

- Prediction: All else being equal, speakers will choose first-last name pairs that are more rhythmically well-formed.

| | | |
|---|---|---|
| e.g., | *SUsan SMITH* | 2172 instances |
| > | *SuZANNE SMITH* | 550 |
| > | *MElanie fitzGErald* | 27 |

## **Factor:** Operationalizing Rhythm

- Eurhythmy Distance (ED): measures how far away from binary alternating rhythm a given construction is. (Shih et al., to appear; cf. Temperley 2009)

  ED = | # of unstressed syllables – 1 |

  *SUsan SMITH* *SuZANNE SMITH*
  |1 – 1| = 0 |0 – 1| = 1

  *MElanie fitzGErald*
  |3 – 1| = 2

## **Controls:** Frequency and Popularity

- Naming choices follow popularity and frequency trends. (Tucker 2001; a.o.)

- Popularity of forename
  - U.S. Social Security Administration: frequencies of 400 most frequent baby names (200 male/200 female) from each decade between 1950 – 2000

- Frequency of surname
  - Frequency of surname in the facebook corpus.

## Corpus studies

1. Frequency of personal name
   a. polysyllabic forenames + monosyllabic surnames
   b. iamb-initial surnames

2. Attested vs. unattested status of personal name

## Study 1a.

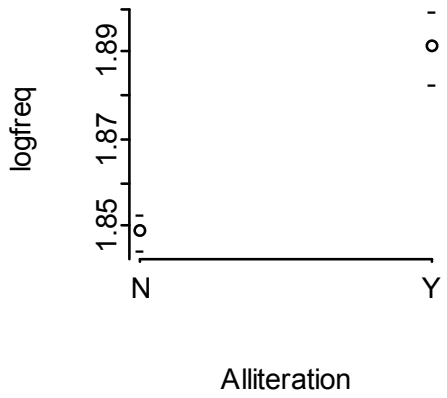
- Polysyllabic forenames + monosyllabic surnames
  $n$ = 806,233 unique personal names

| | | |
|---|---|---|
| most frequent: | *David Smith* | 7440 instances |
| | *Sarah Smith* | 5039 |
| least frequent: | *Donovan Ladd* | 2 |
| | *Dorcus Scott* | 2 |

➢ Prediction: more frequent personal names are more likely to follow phonological preferences.

*Sarah Smith* { alliteration, alternating stress, no OCP violations }

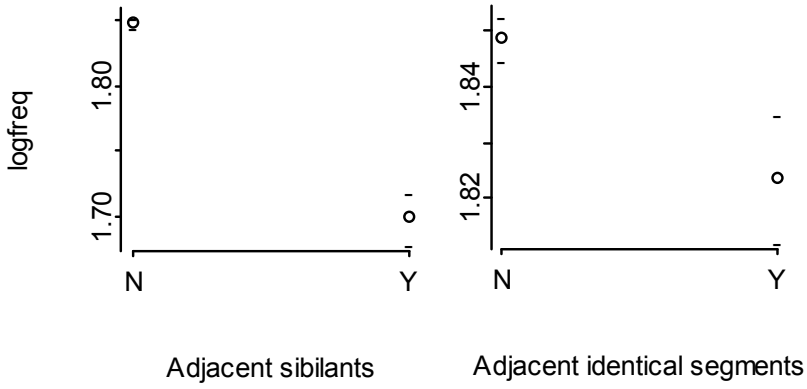
## **Study 1a.** Results




- Frequent name pairs are more likely to be alliterative.

Partial effects – all other predictors held constant.

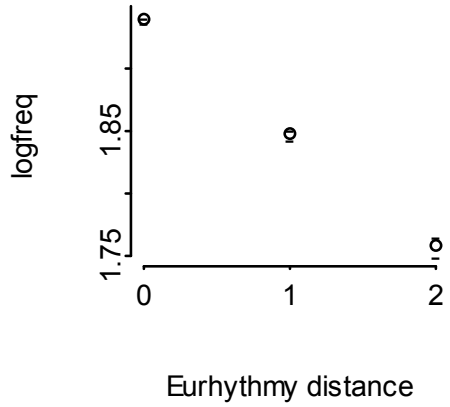
## **Study 1a.** Results




Partial effects – all other predictors held constant.

- Frequent name pairs are more likely to avoid adjacent sibilants and identical segments.

## **Study 1a.** Results

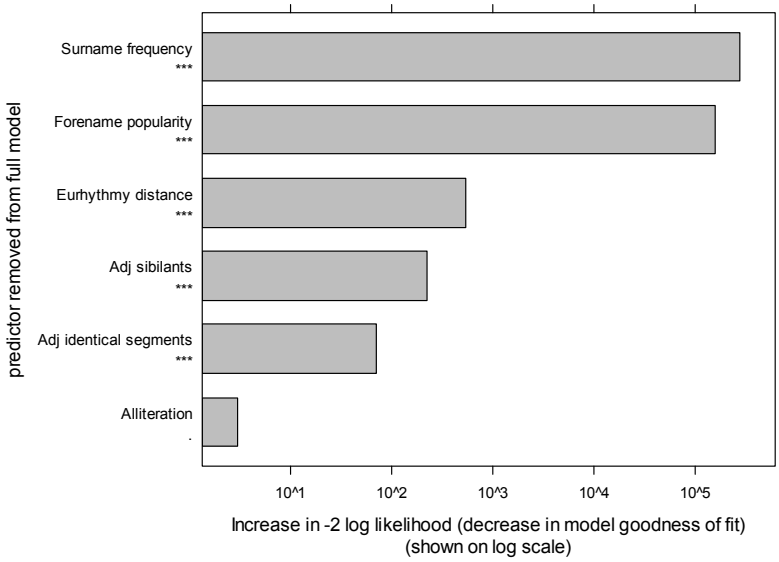



- Frequent name pairs are more likely to exhibit binary rhythm, avoiding clash and lapse.

Partial effects – all other predictors held constant.

**Study 1a.** Results

## Individual Contributions of Predictors



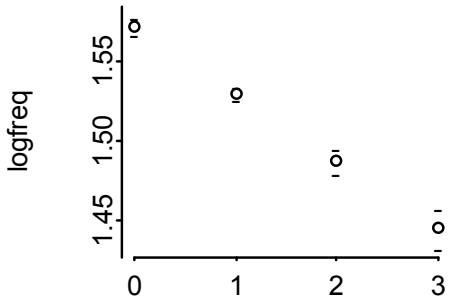
## Study 1b.

- The phonological make-up of English forenames and surnames predisposes pairs to perfect rhythmic patterning.
  - 74.98% of polysyllabic forenames end with a trochee
    - e.g., *DAvid*
  - 77.7% of polysyllabic surnames begin with a trochee
    - e.g., *JOHNson*
  - 26.96% of surnames are monosyllabic

- Iamb-initial polysyllabic surnames (*n* = 286,042)
  - e.g., *Buchanan, Burnett, Fontaine, Levine, Maloney, Marie, McDonald, Montgomery, Munro*

➢ Prediction: Iamb-initial last names should be more frequently paired with stress-final or monosyllabic first names.
  - e.g., *suZANNE fitzGErald > SUsan fitzGErald*
    *SUE fitzGErald >*

## Study 1b. Results




➢ With iambic surnames, frequent name pairs are still more likely to exhibit binary rhythm, avoiding clash and lapse.

Partial effects – all other predictors held constant.

## **Study 1.** Results

- Personal names that are more frequent are more likely to conform to phonological well-formedness preferences.

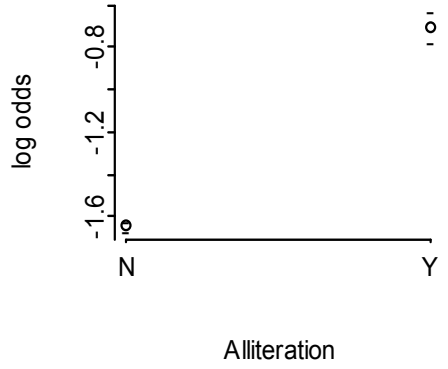
- Given the range of possible personal names, do speakers choose name pairs that better fit with these linguistic preferences over ones that do not?

## **Study 2.** Choosing optimal names

➢ Prediction: Given the range of all possible name pairs, attested personal names are better phonologically formed than name pairs that do not occur.

- Forming a baseline
  - forenames and surnames in corpus were randomly shuffled and checked against the attested name pairs to generate personal names that do not occur.
- Polysyllabic forenames and surnames
  *n* = 3,461,906
  - attested = 1,649,342
  - unattested (generated) = 1,812,564
- Results reported from representative subset (*n* = 300,000)

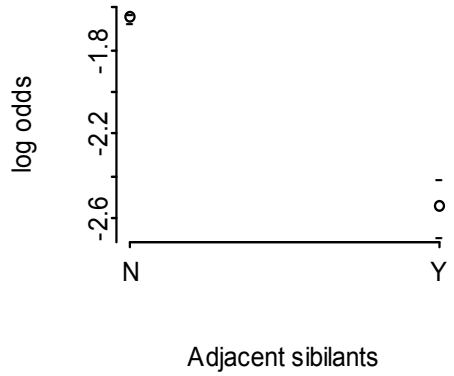
## **Study 2.** Results




- Actual name pairs are more likely to alliterate than names that do not occur.

Partial effects – all other predictors held constant.
Higher log odds value = greater likelihood of attested name

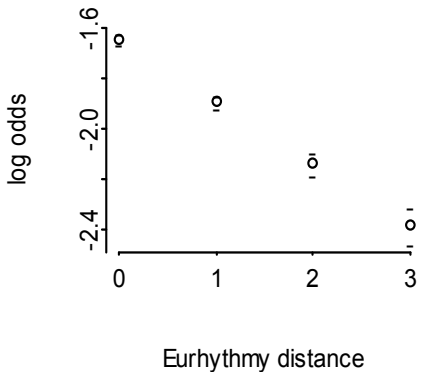
## **Study 2.** Results




- Actual name pairs are more likely to avoid adjacent sibilants.
- Adjacent identical segments (OCP) did not reach significance in this data.

Partial effects – all other predictors held constant.
Higher log odds value = greater likelihood of attested name

## **Study 2.** Results

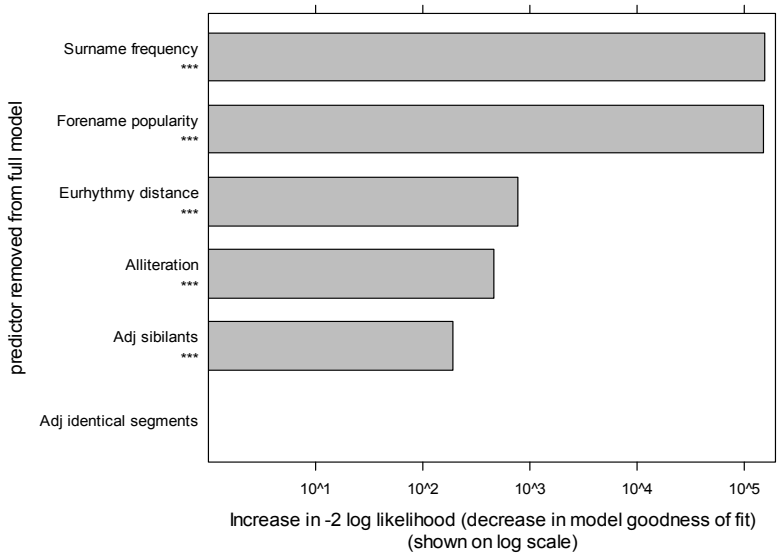



- Actual name pairs are more likely to have binary rhythm, avoiding lapse and clash, than names that do not occur.

Partial effects – all other predictors held constant.
Higher log odds value = greater likelihood of attested name

**Study 2.** Results

## Individual Contributions of Predictors



## **Study 2.** Results

- Attested personal names follow phonological preferences more than other possible combinations of forenames and surnames.

## Discussion

- Controlling for available external factors, phonological preferences affect personal name choice:

In particular,

- rhythmic well-formedness preferences
- avoidance of adjacent sibilants

also,

- alliteration
- avoidance of adjacent identical segments

## Discussion

Avoidance of adjacent identical segments

- Does not distinguish between possible biases to avoid certain clusters or between classes of similar sounds (e.g., sibilants).

  e.g., *Tom Monroe* > *Carl Rogers*
  [m] – [m] [l] – [r]

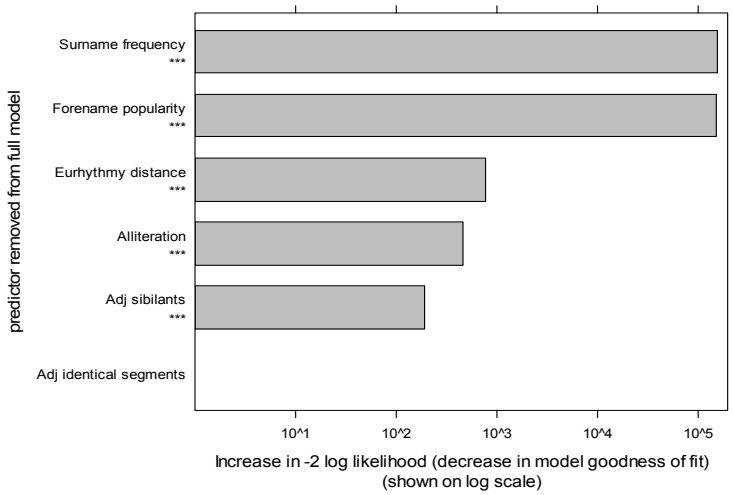
(Martin 2007)

## Discussion

- The phonological preferences used in personal name choice are the same as the ones active in other linguistic processes.
- Observed relative effect sizes between phonological factors and other factors (e.g., frequency, popularity) in name choice are similar to those observed in word and construction choice studies.

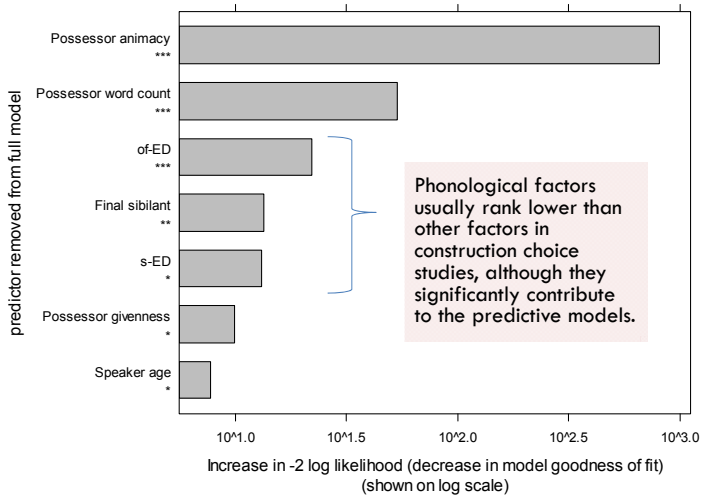
# Discussion

Predictor contributions in personal name choice (Study 2)




# Discussion

Factors in English genitive construction choice (Shih et al., to appear; Grafmiller and Shih, in prep)




Phonological factors usually rank lower than other factors in construction choice studies, although they significantly contribute to the predictive models.

## Discussion

- In addition to sharing phonological factors, similar predictor rankings suggest that the importance of these factors in the overall linguistic system is the same across personal name choice and other linguistic choices.

## **Discussion:** future directions

- Amount of variance explained by the models in Study 1 is fairly low.
  - a. adjusted $R^2 = 0.393$
    b. adjusted $R^2 = 0.265$
  - Corpus limitations: does not incorporate many known social, cultural, and other factors.
- Other determinants
  - rhyme avoidance (e.g., *Joe Monroe*)
  - orthographic alliteration
  - phonotactic and syllable structure preferences
  - information theory (Ramscar et al. 2011)

## Conclusion

- Large scale study of personal names using public access, social media data
- Personal names should be studied as a unit.

➢ When speakers choose names, they access the same internal phonological preferences that drive other linguistic choices, in addition to external linguistic factors, making the study of names a valuable testing ground for investigating such effects.

## *Thank you!*

Acknowledgements to Sharon Inkelas, Arto Anttila, Jason Grafmiller, Joan Bresnan, Rebecca Starr, Roey Gafter, Tom Wasow, Beth Levin, Kie Zuraw, and audiences at Stanford for valuable discussion.

contact: stephsus@stanford.edu
slides: http://www.stanford.edu/~stephsus/shih-LSA2012.pdf

# Select references


Adams, Matthew. 2011. Poetic correspondence and the Welsh *cynghanedd*. Paper presented at the 85th Annual Meeting of the Linguistic Society of America. Pittsburgh, PA. 6–9 Jan.

Anttila, Arto; Matthew Adams; and Michael Speriosu. 2010. The role of prosody in the English dative alternation. *Language and Cognitive Processes*. 25(7-9): 946-981.

Barry, III, Herbert. 2010. Racial and Gender Differences in Diversity of First Names. *Names*. 58(1): 47-54.

Benor, Sarah Bunin and Roger Levy. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language*. 82(2): 233-278.

Bloothoft, Gerrit and Loek Groot. 2008. Name Clustering on the Basis of Parental Preferences. *Names*. 56(3): 111-163.

Bloothoft, Gerrit and David Onland. 2011. Socioeconomic Determinants of First Names. *Names*. 59(1): 25-41.

Boers, Frank and Seth Lindstromberg. 2005. Finding ways to make phrase-learning feasible: The mnemonic effect of alliteration. *System*. 33:225-238.

Bowes, Ron. 2010. Facebook usernames. <http://www.skullsecurity.org/blog/2010/return-of-the-facebook-snatchers> accessed 2010.

Feen. 2010. "rules" for naming your baby – just for fun! *Circle of Moms*. accessed 2011.

Fitt, Susan. 2001. Unisyn lexicon. Centre for Speech Technology Research, University of Edinburgh. <www.cstr.ed.ac.uk/projects/unisyn/> accessed 2010.

Grafmiller, Jason and Stephanie Shih. in prep. Weighing in on end weight. Stanford University.

Hinrichs, Lars and Benedikt Szmrecsányi. 2007. Recent changes in the function and frequency of standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics*. 11(3): 437-474.

Lindstromberg, Seth and Frank Boers. 2008. The Mnemonic Effect of Noticing Alliteration in Lexical Chunks. *Applied Linguistics*. 29(2): 200-222.

Martin, Andrew. 2007. The Evolving Lexicon. Ph.D. dissertation. University of California, Los Angeles.

McDonald, Janet; Kathryn Bock; and Michael Kelly. 1993. Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*. 25: 188-230.

Menn, Lise and Brian MacWhinney. 1984. The repeated morph constraint: Toward an explanation. *Language*. 19: 519-541.

Ramscar, Michael; Asha Halima Smith; Richard Futrell; and Melody Dye. 2011. The distribution of proper names across cultures: How social engineering has undermined an evolved social technology. MS. Stanford University.

Rose, Sharon and Rachel Walker. 2004. A Typology of Consonant Agreement as Correspondence. *Language*. 80(3): 475-531.

Shih, Stephanie. 2011. Unisyn supplement, v.1.

Shih, Stephanie; Jason Grafmiller; Richard Futrell; and Joan Bresnan. to appear. Rhythm's role in English genitive construction choice. In Vogel, R. and R. Van de Vijver (ed). *Rhythm in phonetics, grammar, and cognition*.

Temperley, David. 2009. Distributional Stress Regularity: A Corpus Study. *Journal of Psycholinguistic Research*. 38(1): 75-92.

Tucker, D.K. 2001. Distribution of Forenames, Surnames, and Forename-Surname Pairs in the United States. *Names*. 49(2): 69-96.

Social Security Administration. 2010. Popular Baby Names by Decade. <www.ssa.gov/oact/babynames/decades/index.html> accessed Nov 2010.

Wattenberg, Laura. 2005. *The Baby Name Wizard: A Magical Method for Finding the Perfect Name for Your Baby*. NY: Broadway Books.

Whissell, Cynthia. 2001. Sound and Emotion in Given Names. *Names*. 49(2): 97-120.

Wilen, Joan and Lydia Wilen. 1993. *The Perfect Name for the Perfect Baby*. Ballantine Books.

Wright, Saundra K.; Jennifer Hay; and Tessa Bent. 2005. Ladies first? Phonology, frequency, and the naming conspiracy. *Linguistics*. 43(3): 531-561.

Zuraw, Kie. 2002. Aggressive Reduplication. *Phonology*. 19: 395-439.

Zwicky, Arnold. 1987. Suppressing the Z's. *Journal of Linguistics*. 23: 133-148.